**Capítulo II.**

**2. MARCO TEÓRICO.**

* 1. **Importancia de los datos**

Para dar sentido a la recolección de datos, es decir, para poder obtener información relevante a partir de los datos, debemos agrupar los datos unos con otros de manera que estos tengan diferencias significativas que los caractericen tanto dentro como fuera de cada uno de sus grupos. Sólo así podremos llegar a obtener conclusiones lógicas que nos lleven a la solución de un problema o simplemente para conocer lo que estos datos significan o representan.

Según Levin y Rubin[[1]](#footnote-2) “es necesario tomar en cuenta que todos los grupos relevantes deben estar representados en los datos”, es decir que se debe tratar en lo posible de no omitir grupos de datos que sean representativos para nuestro estudio, por ejemplo, si estamos estudiando la preferencia de los hombres ecuatorianos de acuerdo a su raza a una cierta marca de perfume y los clasificamos simplemente en hombres blancos y hombres negros,

estaremos omitiendo al grupo de los mulatos y de los indios de tal manera que al realizar el análisis correspondiente estaremos obteniendo resultados incompletos.

También “debemos asegurarnos de trabajar con datos confiables” ya que los resultados de los análisis estadísticos que se realicen van a depender única y directamente de los datos colectados. Además “los datos deben estar basados en suposiciones e interpretaciones correctas”. Debemos de poder reconocer el significado de cada una de las variables de manera que no existan contradicciones, ni dudas, ya que de eso dependen las conclusiones y el análisis.

* 1. **Histogramas de frecuencias**

**2.2.1 Ordenando los Datos**

Es muy común encontrarse con muestras de datos en desorden, el primer paso a seguir para graficar un histograma de frecuencias es ordenar los datos. Por ejemplo, a continuación se presenta una tabla de datos que contiene datos de promedios de calificaciones.

16 15 17 08 12

20 16 17 14 18

13 16 16 18 15

17 11 17 16 15

**Tabla 2.1**

Promedios de calificaciones de Matemáticas de estudiantes de tercer año de secundaria.

Procedemos ahora a ordenar los datos en forma ascendente

**Tabla 2.2**

08 11 12 13 14

15 15 15 16 16

16 16 16 17 17

17 17 18 18 20

Promedios de calificaciones de estudiantes de tercer año de secundaria ordenados en forma ascendente.

**2.2.1.1 Ventajas del ordenamiento de datos**

Al arreglar los datos podemos notar características que antes eran difíciles notar a simple vista:

**Podemos notar inmediatamente la cota superior (máximo) y la cota inferior (mínimo).-** En la tabla 2.2 vemos que la mínima nota es 08 y la máxima es 20.

**Podemos notar si existen datos que se repiten**.-Del ejemplo anterior (tabla 2.2) podemos observar que los números 15, 16, 17 y 18 se repiten 3, 5, 4 y 2 veces respectivamente

**Podemos dividir los datos en intervalos.-** De la tabla 2.2 notamos que en la primera mitad se encuentran las notas desde 08 hasta 16, mientras que en la segunda mitad las notas van desde 16 hasta 20, esto nos hace notar que más de la mitad de los estudiantes aprobaron Matemáticas ese año.

**2.2.2 Distribución de Frecuencias**

Una tabla de frecuencias contiene información acerca de la distribución de frecuencias de una variable. Para construir la tabla de frecuencias de un conjunto de datos primero dividimos los datos en intervalos (o clases, no necesariamente del mismo tamaño) que contienen el conjunto de datos de la muestra que cae en ese intervalo, de tal manera que cada uno de los datos estén contenidos en uno y solamente uno de los intervalos. Utilizaremos la tabla 2.2 para generar una tabla de frecuencias en base a los promedios de calificaciones del ejemplo anterior.

| **Clase** | **Frecuencia** |
| --- | --- |
| 05-08 | 1 |
| 09-12 | 2 |
| 13-16 | 10 |
| 17-20 | 7 |

**Tabla 2.3**

Distribución de frecuencias de los promedios de calificaciones de Matemáticas de estudiantes de tercer año de secundaria.

**2.2.3 Histogramas**

Se puede graficar histogramas en el plano cartesiano, en el espacio euclidiano, y en cualquier espacio de dimensión finita, pero es preferible y para cuestiones prácticas graficarlo en **R2** y **R3**, en este texto nos referiremos únicamente a histogramas graficados en el plano cartesiano.

Un histograma de frecuencias consiste en un conjunto de rectángulos colocados verticalmente sobre el eje de las abscisas (eje de las clases: eje **X**) cuyo alto corresponde a la frecuencia observada dentro de cada clase (como se vio anteriormente uno para cada clase: eje **Y**). Si los intervalos que describen a las clases son todos del mismo tamaño el ancho de cada uno de los rectángulos es el mismo, en caso contrario si los intervalos son de diferente tamaño el ancho será proporcional al tamaño de la clase al que corresponda el rectángulo. La figura 2.1 ilustra un histograma de frecuencias para los datos de la tabla 2.2 dividiéndolo en las clases descritas en la tabla 2.3



Histograma de Frecuencias

#### Figura 2.1

Histograma de frecuencias de los promedios de calificaciones de Matemáticas de estudiantes de tercer año de secundaria

**2.3 Probabilidad y Variables Aleatorias.**

**2.3.1 Probabilidad**

Analicemos un ejemplo para entender el concepto. Si hacemos un experimento que consiste en lanzar una moneda al aire una vez, todos los posibles resultados son dos, cara y sello, pero si la lanzamos dos veces el número de resultados posibles son 4 como lo muestran la figura 2.2 y la tabla 2.4:

**1**

**2**

**3**

**4**

cara

sello

cara

sello

cara

sello

**Figura 2.2.**

Diagrama de tallos y hojas que resume los 4 posibles resultados de lanzar una moneda dos veces

|  |  |  |
| --- | --- | --- |
| **1** | Cara | cara |
| **2** | Cara | sello |
| **3** | sello | cara |
| **4** | sello | sello |

#### Tabla 2.4.

#### resume los 4 posibles resultados de lanzar una moneda dos veces.

A partir de estas figuras podemos notar que el número de posibilidades de obtener dos caras de estos lanzamientos es 1 de 4, lo mismo sucede para dos sellos, pero que el número de posibilidades de obtener una cara y un sello es 2 de 4. Podemos resumir esto en la tabla 2.5

|  |  |
| --- | --- |
| cara y cara | 1 de 4 |
| cara y sello ó  sello y cara | 2 de 4 |
| sello y sello | 1 de 4 |

#### Tabla 2.5.

#### resume las posibilidades de obtener cada uno de los resultados después lanzar una moneda dos veces.

En términos matemáticos podemos decir que 1 de 4 es y que 2 de 4 son  , estas dos fracciones definen dos probabilidades, la primera es la probabilidad de obtener dos caras y la segunda es la probabilidad de obtener una cara y un sello en dos lanzamientos de una moneda.

Para entender la definición de probabilidad, definamos primero “espacio muestral”.

*“Un espacio muestral es el conjunto de todos los resultados posibles de un experimento”*

Probabilidad es una función que asigna a cada evento del espacio muestral un número entre cero y uno.

**p:** S [0,1]

x **p(x)**,

donde S es un espacio muestral y x es un evento de S y p(x) que se define como la probabilidad de que ocurra el evento x.

Una función de probabilidad debe cumplir con los siguientes axiomas.

1. p(S)=1
2. 0≤p(x)≤1
3. 

**2.3.2 Variables Aleatorias**

Una variable aleatoria x es una función definida sobre un espacio muestral S, donde, a cada elemento a de S (llamado evento), se le asigna un número real **x**(a).

**x:** S ℜ

a **x**(a), donde a∈S

Así, retomando el ejemplo anterior del lanzamiento de monedas, se puede definir una variable aleatoria **x** como el número de caras que se observan en dos lanzamientos de una moneda, de donde se sabe que **x** puede tomar los valores de 2, 1 ó 0. La tabla de probabilidades para cada uno de estos sucesos se muestra a continuación.

Sea

**x**: número de caras que se observan en dos lanzamientos de una moneda

|  |  |  |
| --- | --- | --- |
| Evento | x | **p**(**x**=x) |
| cara y cara | 2 |  |
| cara y sello ó  sello y cara | 1 |  |
| sello y sello | 0 |  |

#### Tabla 2.6

#### resume las probabilidades de obtener cada uno de los resultados posibles después lanzar una moneda dos veces.

Como se observa la columna **p**(**x**=x) muestra las probabilidades de que la variable aleatoria **x** tome los valores 2, 1 y 0 que son ,  y  respectivamente.

Existen dos tipos de variables aleatorias discretas y continuas

* **Variables aleatorias discretas**

Este tipo de variables aleatorias se caracterizan porque su dominio en un conjunto finito o infinito contable de valores, en otras palabras, estas variables aleatorias pueden tomar un número limitado de valores.

* **Variables aleatorias continuas**

Las variables aleatorias continuas se definen en un dominio continuo o sea pueden tomar un valor cualquiera dentro de un intervalo dado.

Ejemplos de variables aleatorias discretas y continuas:

* **Discretas**

**x**: grado escolar de un niño de primaria

**dom**(**x**) = **⎨**1**,** 2**,** 3**,** 4**,** 5**,** 6**⎬**

**y**: número de páginas de un libro

**dom**(**y**) = **⎨**1**,** 2**,** 3**,** …**,** n**,** …**⎬**

* **continuas**

**w**: peso en kilogramos de una persona

**dom**(**w**) = **(**0.5 **,** 1000**)**

**z**: número real

**dom**(**z**) = **(**-∞ **,** +∞**)**

**2.3.3 Valor Esperado.**

Asociado a cada variable aleatoria **x** existe un número llamado valor esperado E[**x**] que como su nombre lo define es el valor que se esperaría obtener al seleccionar aleatoriamente un valor de los que puede tomar una variable aleatoria. Matemáticamente este valor es calculado dependiendo del tipo de variable que se estudie, si es discreta o continua. Las fórmulas matemáticas para obtener este valor se presentarán en la sección 2.4.

**2.3.4 Densidades de probabilidad.**

Acerca de este tema nos limitaremos a decir que así como hicimos con el ejemplo de las monedas, se puede definir una función de probabilidad para cada variable aleatoria. Esta función se llama función de densidad de probabilidad de la variable aleatoria o simplemente función de distribución de **x** ó f(**x**)=**p**(**x**=x).

**2.4** **Medidas de Tendencia Central y Dispersión.**

Las medidas de tendencia muestran mediante cálculos matemáticos algunas características de la población.

Las medidas de tendencia central nos dan una noción del valor hacia donde se agrupan la mayoría de los valores que toma la variable de estudio. Estos valores se calculan por medio de parámetros poblacionales (en el caso de estudiar poblaciones), o estimadores de estos parámetros (en el caso de trabajar con muestras).

**2.4.1 La Media.**

Supongamos que estudiamos una población cualquiera, se puede calcular la tendencia central de los datos por medio de la media poblacional μ, este valor es un parámetro, es decir una constante, y se calcula mediante las fórmulas.

##### Para variables aleatorias discretas.

,

donde  es el valor esperado de x ([[2]](#footnote-3)) y  es la probabilidad de que la variable aleatoria **X** tome el valor xi y N es el tamaño de la población.

Para variables aleatorias continuas.

,

donde es la función de distribución de x.

Para el caso de querer estimar la tendencia central de una población por medio de una muestra, podemos hacerlo calculando un estimador de la media poblacional, examinemos el caso de  (equis barra), la media aritmética, cuya fórmula se deriva de la anterior.

,

donde xi representa cada uno de los valores que toma la muestra y n es el tamaño de la muestra.

Para ilustrar mejor el significado de μ y  observemos la figura 2.3

#### Figura 2.3

**f(x)**

Ilustración de la media poblacional

μ

**x**

Tenemos aquí una población **X** y hemos graficado su distribución de puntos f(**X**). Podemos notar que los valores que toma **X** se agrupan alrededor de μ, teniendo la máxima frecuencia en μ y desplazándose hacia ambos lados de este valor. Esto es debido a que la distribución de la población es simétrica. Para distribuciones asimétricas la media se mueve del valor con frecuencia máxima hacia la derecha o izquierda, dependiendo del caso.

Si dibujamos la distribución de otra población **X’** en el mismo eje como en la figura 2.4 podemos comparar que las dos poblaciones tienen distinta media y que los valores de **X’** se agrupan hacia la derecha de los valores de **X**, es decir que la

media de **X’** es mayor que la de **X**, esto es obvio dado que μ2>μ1.

#### Figura 2.4

**f(x’)**

**f(x)**

Comparación entre

medias

μ2

μ1

**2.4.2 La Moda**.

Otra manera de medir la tendencia central es mediante el cálculo de la moda. La moda se define como el valor que más se repite entre el conjunto de datos. Es decir para calcular su valor se debe contar cuantos números corresponden a cada dato y el dato con mayor frecuencia de repeticiones es la moda. También podemos dividir los datos en clases y contar los números que corresponden a dicha clase, así sabremos cual es la *“clase modal”*. Examinemos un ejemplo ilustrativo. La tabla 2.7 contiene el tiempo que tardó una cajera de un banco en atender a cada uno de los clientes que llegaron a su ventanilla durante una jornada de trabajo de 2 horas.

**Tabla 2.7**

3 6 3 3 4

4 5 4 5 4

5 4 4 5 3

3 4 6 4 4

5 3 4 4

Tiempo de atención (en minutos) a clientes en una ventanilla de un banco durante dos horas.

Al ordenar y contar los datos obtenemos la tabla 2.8

| **Minutos** | **Frecuencia** |
| --- | --- |
| 3 | 6 |
| 4 | 11 |
| 5 | 5 |
| 6 | 2 |

**Tabla 2.8**

Distribución de frecuencias del tiempo de atención a clientes en una ventanilla de un banco durante dos horas.

#### Figura 2.5

Histograma de frecuencias del tiempo de atención a clientes en una ventanilla de un banco durante dos horas.

De aquí obtenemos que la moda es 4 ya que es el valor que más se repite.

Cuando en una distribución de frecuencias se encuentra más de una moda, esta distribución se llama bimodal, trimodal, etc, dependiendo del número de modas que contenga.

**2.4.3 La Mediana**.

El valor de la mediana  depende directamente del número de datos que contenga la población. Para calcular la mediana primero debemos de ordenar los datos, luego calculamos la posición en la que se encuentra la mediana dependiendo del tamaño de la población (o de la muestra), si es par o impar, y al final calculamos su valor. Dicho en otras palabras, si el tamaño de la población es impar se encontrará en la posición (n+1)/2 y si es par la mediana se define como es promedio entre los valores que se encuentran en las posiciones n/2 y (n+2)/2. Es decir,



**2.4.4 La Varianza**

La varianza (σ2 o s2) es una medida de dispersión, es decir, indica que tan separados (o dispersos) se encuentran los datos. Este valor se calcula obteniendo la suma del cuadrado de las diferencias entre cada uno de los datos y la media y dividiendo este valor para N (el tamaño de la población) o para n-1 (n es el tamaño de la muestra). Es decir,

**Varianza poblacional.**



**Varianza muestral.**



A partir de la varianza se puede calcular el valor de la desviación estándar (σ o s) que se obtiene tomando la raíz cuadrada de la primera.

****

La desviación estándar nos da una noción de que tan dispersos de la media están los datos. De estudios realizados se sabe que por lo menos el 75% de los valores caen entre μ+2σ y μ-2σ y que por lo menos 89% de los valores caen entre μ+3σ y μ-3σ. Veremos más adelante que lo aquí expresado es muy útil, ya que a partir de una muestra podemos calcular la media y la desviación estándar para obtener un intervalos (con un cierto grado de confianza) que contengan un gran grupo de datos dentro de ellos y así estimar poblaciones, sin necesidad de analizar todos los datos de la población.

**2.5 Pruebas de Hipótesis.**

**2.5.1 Definición**

Una hipótesis estadística es un supuesto acerca del valor de un parámetro poblacional en base a una muestra. Cuando se quiere probar o contradecir una hipótesis estadística se realizan lo que se llaman “Pruebas de Hipótesis”, mediante estas pruebas se compara una característica de una variable aleatoria con una constante ó con la misma característica de otra variable aleatoria, tomando en cuenta un cierto grado de error llamado confianza (que es elegido por el investigador),. Se pueden formular infinidades de pruebas de hipótesis, entre las más usadas se encuentran las siguientes:

* Pruebas comparando un parámetro poblacional con un valor en referencia.
* Pruebas acerca de la distribución de probabilidad de una variable aleatoria.
* Pruebas comparando el valor de un parámetro poblacional de una variable aleatoria con los de otra.
* Pruebas acerca de la independencia entre dos variables aleatorias.

Las pruebas de hipótesis estadísticas se formulan postulando un supuesto, versus la negación del supuesto. El supuesto se llama hipótesis nula y se denota por Ho, y la negación se denomina hipótesis alterna y se denota por H1, Observemos los siguientes contrastes acerca la media de una población.

**1. Ho:** = 4

**Vs.**

**H1:** ≠ 4

**2. Ho:** ≥ 4

**Vs.**

**H1:** < 4

El primer contraste de hipótesis postula que la media es igual a 4 versus la media no es igual a 4, el segundo postula que la media es mayor o igual a 4 versus la media es menor que 4. En ambos casos, para poder elegir entre la hipótesis nula y la alterna se deben calcular dos valores, compararlos y de acuerdo al resultado aceptar o rechazar la hipótesis nula.

**2.5.2 La confianza**

Se llama nivel de confianza al grado de error que se puede permitir que suceda en la estimación, como ya se mencionó, este error es manejable (elegido por el investigador). El nivel de confianza es un número real contenido en el intervalo [1,100] y se mide en porcentaje. Podremos hablar entonces de 95% de confianza. El nivel de confianza se relaciona con el nivel de significancia denotado por la letra griega , que representa la probabilidad de rechazar la hipótesis nula cuando es falsa o probabilidad de cometer error tipo 1.

Una vez elegido el nivel de confianza c%, se obtiene  mediante el siguiente cálculo:



Entonces, podemos dejar fija la probabilidad de cometer error tipo 1, o sea la probabilidad de concluir que Ho es verdadero cuando en realidad es falso, y así tener control sobre el error de la estimación.

**2.5.3 Estadístico de prueba, Región de Rechazo y Valor p.**

Para decidir entre aceptar o rechazar la hipótesis nula, se debe calcular un valor  en base a una función matemática. Una vez obtenido este valor se puede optar por dos métodos:

1.  se compara con el ***estadístico de prueba*** θ0 que es el valor que toma una variable aleatoria θ, definida con características de acuerdo a la confianza, a los datos y a suposiciones acerca de la población.

Dependiendo del resultado de esta comparación (= θ0 ó< θ0 ó > θ0) se acepta o rechaza la hipótesis nula.

1. Se obtiene el ***valor p*** de la prueba mediante el cálculo de la fórmula:

, y se toma la decisión de

aceptar o rechazar Ho.

El valor p tiene un uso significativo en la estadística, ya que a partir de este valor y un nivel de significancia dado se puede decidir si aceptar o rechazar la hipótesis nula. Por ejemplo, a 95% de confianza (= 0.05) se acepta la hipótesis nula si p ≥ 0.1; se la rechaza si p ≤ 0.05; y cuando 0.1 < p < 0.05 se dice que hay un región de indecisión (no se puede concluir algo), y no se puede tomar una decisión únicamente en base a la información obtenida, para mejorar la región, se debe optar por aumentar el número de unidades muestreadas, o cambiar el nivel de significancia.

El valor p es la probabilidad de que la variable aleatoria θ tome un valor mayor a ; esto no es más que el área bajo la curva de la función de distribución de probabilidad de la variable aleatoria θ. Gráficamente el valor p de la prueba define una región llamada ***región de rechazo***, esto es debido a que dependiendo de la magnitud de esta región se acepta o rechaza la hipótesis nula.

Por ejemplo, asumamos que θ es Normal con media 0 y varianza 1 (normal estándar).

#### Figura 2.6

Grafica la función de distribución de una variable aleatoria Normal Estándar



**z**

**θ0**

p

región de rechazo

Supongamos que el valor del ***estadístico de prueba*** es θ0 (valor obtenido a partir de la tabla de la distribución Z), el área bajo la curva de z cuado z ≥ θ0 es la llamada ***región de rechazo*** que define el ***valor p*** de la prueba, y para este caso se obtiene mediante el cálculo de la fórmula.

,

donde z es una variable aleatoria normal estándar y f(z) es la función de distribución de probabilidad de z.

Supongamos que p = 0.0001; debido a que p<0.05 se dice que p es pequeño y se puede rechazar la hipótesis nula, según como se le halla definido. A continuación definiremos como se establecen los contrastes de hipótesis.

**2.5.4 Pruebas con parámetros poblacionales**

Para estas pruebas, primero se deben hacer supuestos acerca de la distribución de probabilidad de la variable aleatoria, para luego calcular el valor del estadístico de prueba en base a información obtenida de la muestra.

**2.5.4.1 Pruebas de Medias con Varianza Conocida**

Si la población es distribuida Normalmente o el tamaño de la muestra es lo suficientemente grande para asumir normalidad (n ≥ 30) y se conoce la varianza, la región de rechazo viene dada de la siguiente forma:

|  |  |
| --- | --- |
| **Prueba** | **Rechace Ho en favor de H1 si** |
| Ho: μ = μ0 Vs. H1: μ ≠ μ0 |  |
| Ho: μ ≤ μ0 Vs. H1: μ > μ0 |  |
| Ho: μ ≥ μ0 Vs. H1: μ < μ0 |  |

**Tabla 2.9**

Regiones de rechazo para pruebas de hipótesis de medias con varianza conocida, asumiendo normalidad.

Donde el estadístico de prueba es .

**2.5.3.2 Pruebas de Medias con Varianza Desconocida**

Si la población es normal o el tamaño de la muestra es lo suficientemente grande para asumir normalidad y no se conoce la varianza, la región de rechazo viene dada de la siguiente forma:

|  |  |
| --- | --- |
| **Prueba** | **Rechace Ho en favor de H1 si** |
| Ho: μ = μ0 Vs. H1: μ ≠ μ0 |  |
| Ho: μ ≤ μ0 Vs. H1: μ > μ0 |  |
| Ho: μ ≥ μ0 Vs. H1: μ < μ0 |  |

**Tabla 2.10**

Regiones de rechazo para pruebas de hipótesis de medias con varianza desconocida, asumiendo normalidad.

Donde el estadístico de prueba es .

**2.5.3.3 Pruebas de Diferencia de Medias con Varianza Conocida**

Cuando se conoce la varianza y la población es Normal o el tamaño de la muestra es lo suficientemente grande para asumir normalidad (n ≥ 30) la región de rechazo viene dada de la siguiente forma:

|  |  |
| --- | --- |
| **Prueba** | **Rechace Ho en favor de H1 si** |
| Ho: μ1-μ2 = μ0  Vs.  H1: μ1-μ2 ≠ μ0 |  |
| Ho: μ1-μ2 ≤ μ0  Vs.  H1: μ1-μ2 > μ0 |  |
| Ho: μ1-μ2 ≥ μ0  Vs.  H1: μ1-μ2 < μ0 |  |

**Tabla 2.11**

Regiones de rechazo para pruebas de hipótesis de diferencia de medias con varianza conocida, asumiendo normalidad.

Donde el estadístico de prueba es .

**2.5.3.4 Pruebas de Diferencia de Medias con Varianza Desconocida**

Si se asume que ambas poblaciones son distribuidas normalmente o el tamaño de las muestras (n1 y n 2) son lo suficientemente grandes para asumir normalidad y no se conoce la varianza, pero se asume que las varianzas de ambas poblaciones son iguales (σ1 = σ2), la región de rechazo viene dada de la siguiente forma:

|  |  |
| --- | --- |
| **Prueba** | **Rechace Ho en favor de H1 si** |
| Ho: μ1-μ2 = μ0  Vs.  H1: μ1-μ2 ≠ μ0 |  |
| Ho: μ1-μ2 ≤ μ0  Vs.  H1: μ1-μ2 > μ0 |  |
| Ho: μ1-μ2 ≥ μ0  Vs.  H1: μ1-μ2 < μ0 |  |

**Tabla 2.12**

Regiones de rechazo para pruebas de hipótesis de diferencia de medias con varianzas desconocida pero iguales, asumiendo normalidad.

Donde el estadístico de prueba es  y 

Si por el contrario se asume que las varianzas de ambas poblaciones no son iguales (σ1 ≠ σ2), la región de rechazo viene dada de la siguiente forma:

|  |  |
| --- | --- |
| **Prueba** | **Rechace Ho en favor de H1 si** |
| Ho: μ1-μ2 = μ0  Vs.  H1: μ1-μ2 ≠ μ0 |  |
| Ho: μ1-μ2 ≤ μ0  Vs.  H1: μ1-μ2 > μ0 |  |
| Ho: μ1-μ2 ≥ μ0  Vs.  H1: μ1-μ2 < μ0 |  |

**Tabla 2.13**

Regiones de rechazo para pruebas de hipótesis de diferencia de medias con varianzas desconocida y diferentes, asumiendo normalidad.

Donde el estadístico de prueba es  y 

**2.6 Tablas de contingencia.**

**2.6.1 Definición**

Son arreglos de filas y columnas sobre las que se contrastan 2 características medibles de un conjunto de entes o individuos. A cada característica se la puede dividir en un conjunto de criterios mutuamente excluyentes (no necesariamente exhaustivos) que corresponden a los rótulos de las filas y columnas de la tabla (uno para cada criterio).

Tendremos entonces n criterios para la primera característica y p criterios para la segunda, que en la tabla representan n filas y p columnas respectivamente. El valor de cada celda Xij de la tabla se determina mediante el conteo del número de individuos que cumplen con el criterio i de la primera característica y a la vez con el criterio j de la segunda. A continuación se muestra la tabla completa.

**Tabla 2.14.** Muestra del ingreso de datos en una Tabla de Contingencia.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Característica 2** | | | | | | |
|  |  | **Criterio 1** | **Criterio 2** | **…** | **Criterio j** | **…** | **Criterio p** | **Suma de Filas** |
| **Característica 1** | **Criterio 1** | **X11** | **X12** | … | **X1j** | … | **X1p** | **X1.** |
| **Criterio 2** | **X21** | **X22** | … | **X2j** | … | **X2p** | **X2.** |
| **…** | … | … |  | … |  | … | **…** |
| **Criterio i** | **Xi1** | **Xi2** | … | **Xij** | … | **Xip** | **Xi.** |
| **…** | … | … |  | … |  | … | **…** |
| **Criterio n** | **Xn1** | **Xn2** | … | **Xnj** | … | **Xnp** | **Xn.** |
| **Suma de Columnas** | **X.1** | **X.2** | … | **X.j** | .. | **X.p** | **X..** |

donde **Xij** es la cantidad de individuos que cumplen con el criterio i y el criterio j. Nótese que la sumatoria de los valores de cada fila **Xi.** corresponden al número de individuos que cumplen con el criterio i de la característica 1 y la sumatoria de los valores de cada columna **X.j** corresponde al número de individuos que cumplen con el criterio j de la característica 2.

Además, la sumatoria de los valores de todas las filas debe ser igual a la sumatoria de los valores de todas las columnas y a la vez al número total de individuos en estudio; en otras palabras.



Una vez completada la tabla se procede a analizar el contraste de hipótesis siguiente.

**2.6.2 Hipótesis**

**Ho: La característica 1 es independiente de la característica 2**

**Vs.**

**H1: No es verdad Ho([[3]](#footnote-4))**

A (1-α)100% de confianza se rechaza la hipótesis nula Ho a favor de la hipótesis alterna H1 si:

***X2\*>χ2α,(n-1)(p-1)***

donde ,  , , y χ2α,(n-1)(p-1) es el valor correspondiente a una variable aleatoria Ji-cuadrado con (1-n)(1-p) grados de libertad.

**2.6.3 Contribución de las tablas de contingencia al estudio de la sustracción de vehículos en la ciudad de Guayaquil.**

Trataremos de probar que existe relación entre cada par de escenarios bajo los que se han suscitado robos de vehículos, por ejemplo intentaremos probar que existe relación entre la hora y el lugar del robo.

**2.7 Componentes Principales.**

**2.7.1 Introducción al Análisis Factorial**

En ocasiones nos encontramos con casos en los que tenemos que tratar con muchas variables, lo que se puede derivar en complicados análisis de los resultados. Existen técnicas estadísticas para reducir variables sin perder información, una de estas técnicas es el Análisis de Factores o “Análisis Factorial”.

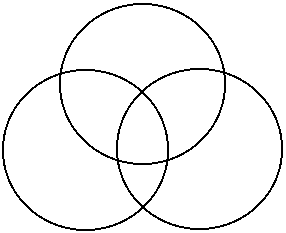
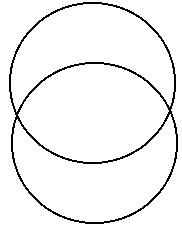
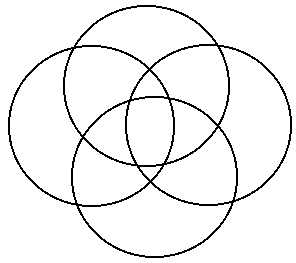
El Análisis Factorial es una técnica multivariada que identifica variables relacionadas y las resume en un conjunto de nuevas variables o factores con el objeto de remover la redundancia y la información duplicada de un conjunto de variables correlacionadas, agrupándolas en algunos factores que contienen variables con características similares dentro del factor y diferentes fuera del factor.

Supongamos por ejemplo que estudiamos 9 variables de interés **X1, X2, X3, X4, X5, X6, X7, X8** y **X9,** el Análisis Factorial dividirá las variables en grupos llamados factores, por ejemplo en 3 factores como lo muestra la figura 2.5

**Figura 2.5**

Ejemplo de variables correlacionadas agrupadas en factores.

**X4**



**Factor 1**

**X1**

**X4**

**X5**

**X8**

**Factor 2**

**X3**

**X7**

###### Factor 3

**X2**

**X6**

**X9**

**X1**

**X5**

**X8**

**X3**

**X7**

**X2**

**X6**

**X9**

v

Cada factor contiene un grupo de variables altamente correlacionadas entre ellas y no correlacionadas entre las demás que no pertenecen a ese factor, permitiéndonos trabajar con menos variables y simplificando el análisis.

Es obvio que en la práctica no se darán casos perfectos como el aquí mostrado en los que las variables de un factor no están correlacionadas con las variables de otro factor, pero se puede esperar que exista poca correlación entre estas variables y mucha entre las variables de un mismo factor.

**2.7.2 Nociones Básicas.**

Componentes Principales es una técnica multivariada de interdependencia que se deriva del Análisis Factorial, en la que cada uno de los factores se expresa como una combinación lineal de las variables (de estudio) que pertenecen a él.

Al aplicar esta técnica al principio se obtienen igual número de factores que de variables, para luego ser reducidos en un número menor de acuerdo a un criterio (de importancia) que nosotros elegimos. Cada factor explica una fracción de la varianza total inherente en el conjunto de datos, así, cada factor tratará de explicar la mayor cantidad de varianza posible. El primer factor siempre contendrá más información que el segundo y más información que la que pueda proporcionar una sola variable, el segundo factor contendrá mas información que el tercero y así en adelante. De tal manera que el o los últimos factores deberán contener individualmente menos información que la que pueda contener una sola variable. Es por esto que podemos eliminar factores, ya que asumimos como poco representativos los que proporcionen menos información que una sola variable.

**2.7.3 La Matriz de Datos**

Primero ingresamos la matriz original de **** que contiene n objetos o entes como rótulos de las filas y p variables o características de los objetos como rótulos de las columnas así como lo muestra la tabla 2.15.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Variables | | | |
| Objetos |  | X1 | X2 | … | Xp |
| O1 |  |  |  |  |
| O2 |  |  |  |  |
| … |  |  |  |  |
| On |  |  |  |  |

Tabla 2.15

Clasificación e ingreso de datos

**2.7.4 El Análisis Preliminar**

El Análisis de Componentes Principales se basa en la matriz de correlaciones o en la de covarianzas, que son calculadas en base a una matriz de datos como la definida en la tabla 2.15, para nuestro estudio nos basaremos en la matriz de Covarianzas. Para obtener esta matriz se calculan las covarianzas σi,j entre cada par de variables Xi y Xj; i,j=1,2,…,p de la siguiente manera:



Con estos valores completamos la matriz de correlaciones que obviamente será cuadrada de ****. Esta matriz se verá como la mostrada en la tabla 2.16

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Variables** | | | | |
| **Variables** | **Xi Xj** | **X1** | **X2** | **…** | **Xp** |
| **X1** | σ11 | σ12 | … | σ1p |
| **X2** | σ21 | σ22 | … | σ2p |
| **…** | … | … |  | … |
| **Xp** | σp1 | σp2 | … | σpp |

Tabla 2.16

Matriz de covarianzas

**2.7.5 Las Componentes Principales.**

Son algebraicamente combinaciones lineales particulares de las p variables aleatorias **X1, X2, X3…Xp**. Geométricamente, estas combinaciones lineales representan la selección de un nuevo sistema coordenado obtenido por la rotación del sistema original con **X1, X2, X3…Xp** como los ejes coordenados iniciales. Los nuevos ejes representan las direcciones con máxima variabilidad y proveen una descripción simple de la estructura de la covarianza.

Como ya se mencionó, los componentes principales dependen de la matriz de covarianza Σ (o de la matriz de correlación ρ) de **X1, X2, X3…Xp**. y su desarrollo no requiere un supuesto normal multivariado.

Sea el vector aleatorio **XT**=[ **X1, X2,…,Xp** ] con matriz de covarianzas Σ y con valores propios λ1≥λ2≥…≥λp≥0.

Considere las combinaciones lineales.

Y1= **a1TX** = a11X1+ a12X2+…+ a1pXp [[4]](#footnote-5)

Y2= **a2TX** = a21X1+ a22X2+…+ a2pXp

…

Yp= **apTX** = ap1X1+ ap2X2+…+ appXp

Y dado que,

Si **Z = cX**, entonces **μz =** Ε[z] = Ε[cX] = **cμx**

y **Σz =** cov[z] = cov[cX] = **cΣxcT**

Entonces,

**var[Yi] =** **aiTΣai**, i = 1,2,..,p

y, **cov[Yi, Yk] = aiTΣak** k = 1,2,..,p

Los componentes principales son aquellas combinaciones lineales no correlacionadas **Y1, Y2,…Yp** cuyas varianzas **var[Yi]** sean lo más grande posible.

La primera componente principal es la combinación lineal **a1TX** que maximice el valor de **var[a1TX ] =** **a1TΣa1**, sujeto a la condición **a1Ta1 = 1**.

La segunda componente principal es la combinación lineal **a2TX** que maximice el valor de **var[a2TX]**, sujeto a las condiciones **a2Ta2 = 1** y **cov[a1TX, a2TX ] = 0**.

Y la i-ésima componente principal es la combinación lineal **aiTX** que maximice el valor de **var[aiTX]**, sujeto a las condiciones **aiTai = 1** y **cov[aiTX, akTX ] = 0** para **k<i**.

Una manera sencilla de encontrar las componentes principales de un conjunto de datos es calculando los valores y vectores propios de su matriz de covarianzas. Sea **Σ** la matriz de covarianzas asociada con el vector aleatorio **XT**=[**X1, X2,…,Xp**], y sean **(λ1,e1), (λ2,e2),…, (λp,ep)**, los pares ordenados valor propio, vector propio de **Σ** donde **λ1≥λ2≥…≥λp≥0**.

La i-ésima componente principal estará dada por:

**Yi =** **eiTX = e1,iX1+ e2,iX2+…+ ep,iXp, i=1,2,…,p**

Además, bajo estas condiciones es claro que;

**var[Yi] =** **eiTΣei**, **= λi,**  i = 1,2,..,p

y, **cov[Yi, Yj] = eiTΣej** **= 0**, i ≠ j; j = 1,2,..,p

**2.7.6 Cálculo de los valores propios.**

A partir de la matriz de covarianzas podremos calcular los valores propios que corresponderán a cada componente obteniendo las soluciones de la ecuación:

;

donde Σpxp es la matriz de covarianzas de las p variables originales, λ son cada uno de los k ≤ p valores propios que vamos a calcular y la matriz Ipxp es la matriz identidad.

Evaluando los valores correspondientes en esta ecuación obtendremos un polinomio de grado menor o igual a p que nos llevarán al cálculo de p valores propios. Al cabo de este paso proseguiremos a ordenar los valores propios de mayor a menor. Como nuestro objetivo es encontrar k<p componentes debemos elegir un criterio de selección.

El valor propio **λi** representa que tanta información contiene la i-ésima componente; es decir, si es mayor que



la i-ésima componente contendrá más información que la que debería proporcionar una variable (en caso contrario menos). Es decir que para nuestro estudio elegiremos el número de componentes a estudiarse de acuerdo al número de valores propios mayores al promedio de los lambdas. Los vectores propios correspondientes a cada valor propio serán los vectores de cada componente principal a estudiarse. Así; tendremos tantas componentes como valores propios mayores al promedio de los lambdas.

Un valor importante es la proporción de información que contiene la l-ésima componente principal. Este valor se calcula dividiendo el l-ésimo valor propio para el total acumulado entre todos los valores propios:

=

% de la varianza total dado por el l-ésimo componente

**2.7.7 La Matriz de Factores.**

En esta matriz la información nos empieza a dar la cara; la matriz de factores se verá de la siguiente manera:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Y1** | **Y2** | **…** | **Yk** |
| **X1** | e11 | e12 | … | e1k |
| **X2** | e21 | e22 | … | e2k |
| **…** | … | … |  | … |
| **Xk** | ep1 | ep2 | … | epk |

Tabla 2.16

Matriz de factores

Esta es una matriz cuyas columnas son los vectores propios correspondientes a los valores propios mayores a 1. cada uno de sus valores **ej,i** son los coeficientes de la i-ésima componente para la j-ésima variable y son directamente proporcionales a las correlaciones entre este componente y esta variable.

Por medio de esta matriz elegiremos en que componente debemos ubicar a cada variable. Mientras mayor sea **ej,i** más se relacionarán **Yi** con **Xj** y viceversa. Por cuestiones prácticas; es usual añadir una columna **** a un lado de cada vector propio (como referencia) para evitar el problema de interpretación de la matriz de covarianzas causado por las diferentes escalas de medición de las variables. En otras palabras ubicaremos a la variable **Xj** en la componente cuyo valor [abs(****)] se aproxime más a 1. La ecuación para el cálculo de las correlaciones entre **Yi** y **Xk** es:



donde: i, k = 1, 2, …, p

1. Richard Levin y David Rubin, autores del libro “Estadística Aplicada para Administradores” sexta edición, año 1996, página12. [↑](#footnote-ref-2)
2. Ver sección 2.3.3 [↑](#footnote-ref-3)
3. H1 también puede ser interpretado como: la característica 1 y la característica 2 son dependientes. [↑](#footnote-ref-4)
4. Denotaremos **X** como el vector **X** [↑](#footnote-ref-5)